Generalizable Semi-supervised Learning Strategies for Multiple Learning Tasks using 1-D Biomedical Signals

Luca Cerny Oliveira [1] **Zhengfeng Lai** [1] **Heather Siefkes** [2] **Chen-Nee Chuah** [1]
[1] Department of Electrical and Computer Engineering, University of California, Davis
[2] Department of Pediatrics, University of California, Davis
{lcernyo,lzhengfe,hsiefkes,chuah}@ucdavis.edu

Abstract

Progress in the sensors field has enabled collection of biomedical signal data, such as photoplethysmography (PPG), electrocardiogram (ECG), and electroencephalogram (EEG), allowing for application of supervised machine learning techniques such as convolutional neural networks (CNN). However, the cost associated with annotating these biomedical signals is high and prevents the widespread use of such techniques. To address the challenges of generating a large labeled dataset, we adapt and apply semi-supervised learning (SSL) frameworks to a new problem setting, i.e., artifact detection in PPG signal and verified its generalizability in ECG and EEG as well. Our proposed framework is able to leverage unlabeled data to achieve similar PPG artifact detection performance obtained by fully supervised learning approach using only 75 labeled samples, or 0.5% of the available labeled data.

1 Introduction

1.1 Background

Congenital Heart Disease (CHD) is the most common birth defect [1]. Critical Congenital Heart Disease (CCHD) is the most severe form of CHD and the leading cause of birth-defect associated infant death. Oxygen saturation (SpO_2) screening is currently a mandated screen test for CCHD. However, despite implementation of mandated screening it is estimated 4.5 CCHD-related early infant deaths occur per 100,000 live-births in the United States [2].

With the challenge of improving CCHD detection in mind, a study developed a dual PPG collection to gather more information from the newborn than traditional pulse oximetry measurement [3]. This data was later used to train a Machine Learning (ML) classifier that outperformed SpO_2 screening [4]. Other work developed an automated motion artifact detection classifier to aid in PPG collection in newborns [5], which outperfomed signal processing methods. Despite these advances in PPG-based CCHD detection, all studies employed supervised learning techniques [4,5]. These approaches require labor-intensive expert annotations, and may introduce incorrect labels if the annotating task is opaque or subjective. Semi-supervised learning (SSL) can combat this issue by leveraging unlabeled data. It also aids with noisy labels, as the SSL framework learns from the signal and not label.

The main contribution of this paper is to adapt and apply SSL framework to a new dataset and learning task: motion artifact detection in PPG signals. Our proposed SSL framework achieves supervised learning performance with 0.5% of labels. We also demonstrate that our SSL framework can generalize to other existing ECG and EEG classification tasks.

Workshop on Learning from Time Series for Health, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

1.2 Related Work

SSL frameworks such as FixMatch [6], FlexMatch [7], Unsupervised Data Augmentation (UDA) [8], MixMatch [9], and Mean Teacher [10] were tested and succeeded in reducing labeling needs on image domain datasets such as MNIST [11], CIFAR [12], and other specialized datasets that include but are not limited to medical imaging datasets [13-15]. Compared to image domain, SSL applied to Time Series Classification (TSC) data has not gained as much traction in the ML research community [16]. The study of generalizable TSC SSL frameworks on biomedical signals showed benefits over supervised learning [17,18]. Jawed *et. al* demonstrated the benefits of general TSC SSL on many datasets including ECG signals [17]. Later, Cheng *et. al* applied their proposed SSL framework to ECG and EEG signals [18].

The rest of the paper is organized as follow. Section 2 describes our dataset, SSL frameworks, and data augmentation strategies employed in our experiments. Section 3 presents the results and Section 4 concludes the paper.

2 Methods

2.1 Problem Setups

This study performs 3-fold cross-validation of our SSL framework with varying percentages of labeled data. We partition our dataset into training (including labeled and unlabeled) sets, and the testing set at the subject level.

PPG This work uses an in-house PPG dataset collected from 472 newborns in their first few days after birth. Each newborn had five minutes PPG waveform collected from hand and foot simultaneously. Newborns spending additional days in the hospital would have repeat measurements taken after each day. There are motion artifact annotations for 99 newborns' PPG waveforms. Two trained observers annotated the dataset. The annotators disagreed on 9.71% of the waveforms' labels. Our learning task was to classify motion artifact. Each pulse was the segment between two consecutive onsets. Every pulse considered artifact by either annotator was labeled as artifact. A total of 1,443,208 pulses were collected, with 7.92% of these pulses being labeled. A total of 49.02% of labels represented artifacts. This PPG dataset is larger than the ones used in prior artifact detection work [5].

EEG This study obtained EEG data from the publicly available Motor Imagery Dataset [19]. A cohort of 109 subjects performed and imagined hand and foot motion tasks while measured by a 64-channel EEG recording device [19]. For our evaluation purpose, we consider the same learning task as described in [20], i.e., binary classification of hand and foot tasks, while ignoring waveforms associated with resting. The entire 4s period of activity is used for classification. We applied a notch filter at 60Hz, and band-pass filter in the 2Hz-60Hz range for noise reduction. Subjects 38, 88, 89, 92, 100, and 104 were removed from the experiment due to incorrect labels [20].

ECG We used the ECG data from the publicly available MIT-BIH Database [21]. Dual-lead ECG recordings were collected from 47 subjects. There are 48 sets of 30-minute long ECG signals collected. The dataset includes annotations for each heartbeat segment's classification. We followed the methodology in [22] and only used the MLII lead of each ECG signal for training and testing. We used R-peak annotations from the metadata to determine heartbeat slices. We removed all patients with paced heartbeats. We followed Association for the Advancement of Medical Instrumentation (AAMI) recommendation of five classes: Non-ectopic (N), Ventricular Ectopic (VEB), Supraventricular Ectopic (SVEB), Fusion (F), and Unknown (Q).

2.2 SSL Frameworks Evaluated

We will consider the following SSL framework where the entire data $\mathcal{D}$ contained a labeled training set $\mathcal{X} = \{(x_i, y_i)\}$ and an unlabeled set $\mathcal{U} = \{x_j\}$. For every data point x_i that is fed to the classifier $g(;\theta)$ by the operation $g(x, \theta) = k$, the maximum predicted confidence element in the vector k is considered the label l_i given by the classifier $g(;\theta)$ to the data point x_i. This operation did not change the parameters θ.

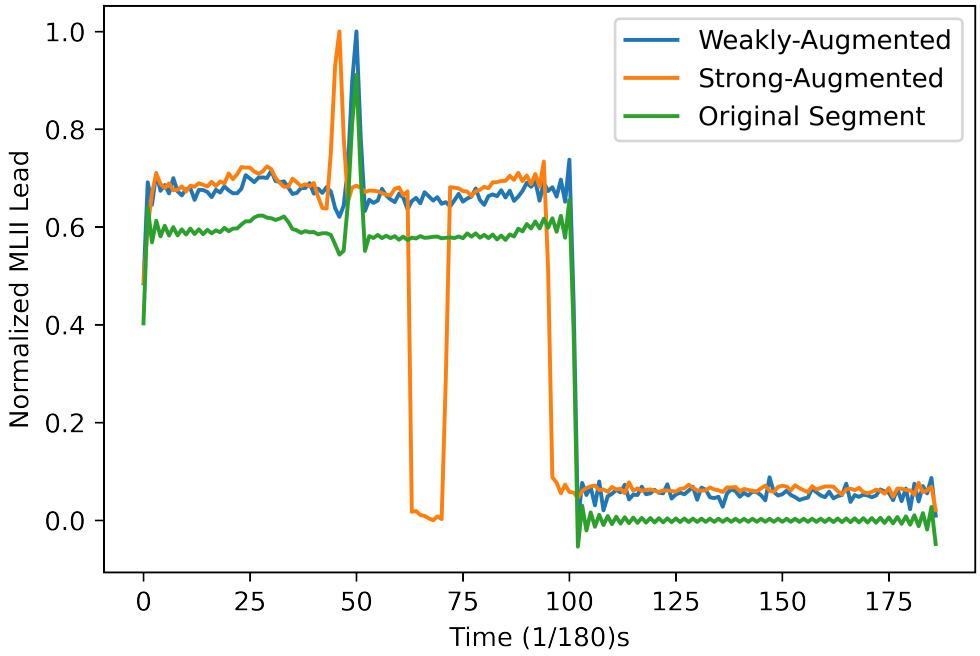

Figure 1: FixMatch and FlexMatch data augmentation strategies applied to ECG heartbeat segment. The weakly-augmented sample had jitter and y-axis offset applied at random intensities. The strongly-augmented sample had all four augmentations (cutout, jitter, x-axis offset, and y-axis offset) applied at random intensities.

Pseudo-labeling Considered to be one of the simplest SSL frameworks, pseudo-labeling, or self-training, uses the classifier's own predictions on unlabeled data points as labels. Hence we can expand the labeled training set $\mathcal{U}$ by using the following operation:

$$\mathcal{X} = \mathcal{X} \cup (\mathcal{P}, L(\mathcal{P})),$$

where $\mathcal{P}$ represents a set of data points from $\mathcal{U}$. $L(\mathcal{P})$ is a set of labels $\{L(x)\}_{x \in \mathcal{U}}$.

FixMatch [6] One of the most popular image-domain SSL framework, FixMatch combines pseudo-labeling with consistency regularization to generate artificial labels during the training process. FixMatch uses cross-entropy loss to "reward" the classifier that outputs the same label for strongly and weakly augmented versions of the same data point, with the condition that prediction confidence is above a 0.95 threshold for the weakly augmented sample.

FlexMatch [7] Similar to FixMatch, FlexMatch also seeks to "reward" models that output consistent predictions for similar inputs through the use of weakly augmented and strongly augmented samples. However, FlexMatch adapts a variable threshold instead of the fixed 0.95 value used by FixMatch. This variable threshold is calculated for each individual class, allowing for underrepresented classes to have more impact in the consistency regularization process through lower thresholds. FlexMatch changes allow for improved performance in multi-class classification [7].

Unsupervised Data Augmentation (UDA) [8] The most general SSL framework tested by our study, UDA has been successfully applied to both image-domain and text-domain. Its generalizability is of interest to our study, as we seek to generalize these SSL frameworks to a different domain: with time-series data. UDA uses consistency loss between the unlabeled sample x_j and a augmented version of the sample $A(x_j)$, where $A(.)$ is a function that employs a random amount of augmentation strategies at random intensities. Therefore, unlike FixMatch and FlexMatch, the augmentations are done completely at random and do not follow a specific weakly augmented or strongly augmented pattern.

2.3 Data Augmentation Strategies

Data augmentation is employed by three out of the four SSL frameworks included in this study. We considered data augmentation strategies that result with least interference in the physiological information embedded in the biomedical signals. In our study, each biomedical signal slice could be augmented using up to four strategie: cutout, jitter, x-axis offset, and y-axis offset. Cutout consists of setting a consecutive slice of datapoints to zero. Jitter involves adding Gaussian modeled noise to the entirety dataset slice. The x-axis and y-axis offset mean displacing the waveform by employing re-sampling and proportional summing. As seen in Figure 1, weak augmentation is consisted of up to two augmentation strategies simultaneously applied to input, while strong augmentation involves applying three or all of the augmentation strategies simultaneously.

3 Results

3.1 Biomedical Signals SSL Proof-of-Concept

PPG When evaluating SSL frameworks on PPG we verified FlexMatch to be the best performing framework. As seen in Figure 2, our study observed that SSL using 0.5% labeled data achieves similar performance as fully supervised models using all the labeled data.

EEG When evaluating SSL frameworks on EEG, UDA as the best performing framework. We observe in Figure 2 that the benefit of applying SSL to EEG. However, the performance of SSL with reduced labeled dataset displays lower accuracy compared to fully supervised models using 100% labeled dataset.

ECG FlexMatch is the best performing framework for ECG learning task. As stated through Figure 2, the scenario with highest improvement from SL to SSL in ECG is achieved at 15% labels.

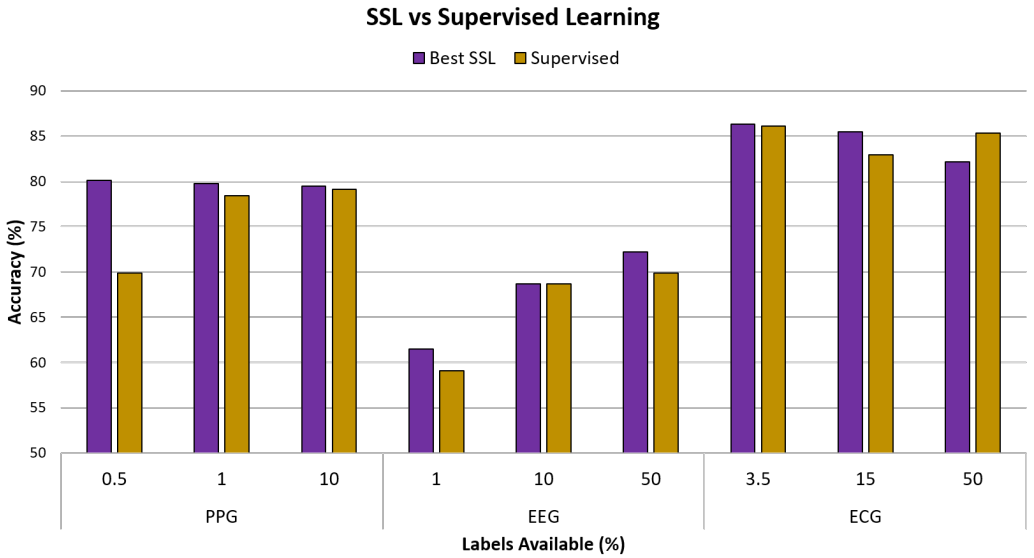

Figure 2: Accuracy plot for all the SSL and SL frameworks experiments using varying amount of labeled data. With 3-fold cross-validation, the 100% labeled scenario is equivalent to employing 67% of labels as training, and 33% of the labels for validation in each fold.

4 Discussion and Conclusions

This study experimented with existing SSL frameworks applied to PPG motion artifact, a relevant classification task in healthcare due to PPG's noise prone nature. We successfully applied the

proposed methods on an in-house PPG artifact detection dataset, where we observed improvement with reduced labeled dataset, with 0.5% of labels.

We evaluated the same proposed frameworks on healthcare-relevant classification tasks using publicly available EEG [19] and ECG datasets [21]. The results deviated from those observed on PPG. While the EEG motor imagery classification benefited from SSL frameworks, the improvement was smaller than seen in PPG. For the ECG dataset evaluation, the benefit of SSL is unclear. We attribute the unclear results in ECG to the overfitting problem caused by data division and delabeling process. There are few samples of SVEB and VEB heartbeats, and these samples are represented in a small number of patients. Moreover, due to AAMI label standard, SVEB and VEB encompass many different arrhythmic heartbeats. In some reduced-label folds, most of the SVEB and VEB samples come from a small number of patients, thus potentially preventing extraction of cross-subject generalizable features for those classes.

There are limitations to our study. First, we used the same data augmentation strategies for all problem setups. Our proposed frameworks were designed to work "off the shelf" for any biomedical signal. We acknowledge that signal-specific data augmentation strategies may increase the performance of UDA, FixMatch, and FlexMatch. For example, Chen *et. al* proposed signal-specific augmentations for EEG [18]. Additionally, our work has the limitation of distribution shift between labeled and unlabeled set. In SSL, there is an assumption that the labeled and unlabeled data share an identical class distribution, which may not be true in real-world scenarios. Since there are no annotations for our in-house dataset, this assumption cannot be guaranteed for our PPG dataset. This assumption is also not met for the MIT-BIH dataset, where the minority classes samples cannot be equally divided between label and unlabeled set for cross-subject analysis.

In our future work, we plan to experiment with additional data augmentation strategies. Additionally, we seek to address distribution shifts between labeled and unlabeled sets. Recent work are successful in improving the performance of SSL frameworks under distribution shift [23], including FixMatch. Applying and building upon these modifications [23] would be beneficial to our framework as biomedical signals' datasets are often imbalanced.

Acknowledgments and Disclosure of Funding

This study was supported by the DataLab at the University of California, Davis, through the Translational Health Data Fellowship program. The project described was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health (NIH), through grant number UL1 TR001860 and linked award KL2 TR001859, the Eunice Kennedy Shriver National Institute of Child Health & Human Development, NIH, through grant number 1R21HD099239-02, the University of California, Davis (UCD) Artificial Intelligence Seed Grant and UCD Venture Catalyst DIAL Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, DataLab, or UCD. The authors would like to thank Diana Guzman, Anushri Parikh, and Palak Pandit, the annotators for the in-house PPG dataset used in this study.

References

[1] Reller, M. D., Strickland, M. J., Riehle-Colarusso, T., Mahle, W. T., & Correa, A. (2008). Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005. *The Journal of pediatrics*, **153**(6), 807-813.

[2] Abouk, R., Grosse, S. D., Ailes, E. C., & Oster, M. E. (2017). Association of US state implementation of newborn screening policies for critical congenital heart disease with early infant cardiac deaths. *Jama*, **318**(21), 2111-2118.

[3] Doshi, K., Rehm, G. B., Vadlaputi, P., Lai, Z., Lakshminrusimha, S., Chuah, C. N., & Siefkes, H. M. (2021). A novel system to collect dual pulse oximetry data for critical congenital heart disease screening research. *Journal of clinical and translational science*, **5**(1).

[4] Lai, Z., Vadlaputi, P., Tancredi, D. J., Garg, M., Koppel, R. I., Goodman, M., ... & Siefkes, H. (2021, November). Enhanced critical congenital cardiac disease screening by combining interpretable machine learning algorithms. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1403-1406). IEEE.

[5] Oliveira, L. C., Lai, Z., Geng, W., Siefkes, H., & Chuah, C. N. (2021, December). A Machine Learning Driven Pipeline for Automated Photoplethysmogram Signal Artifact Detection. In *2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (pp. 149-154). IEEE.

[6] Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33, 596-608.

[7] Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., & Shinozaki, T. (2021). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34, 18408-18419.

[8] Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 6256-6268.

[9] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32.

[10] Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

[11] LeCun, Y. (1998). The MNIST database of handwritten digits. http://yann. lecun. com/exdb/mnist/.

[12] Krizhevsky, A., & Hinton, G. (2010). Convolutional deep belief networks on cifar-10.

[13] Lai, Z., Wang, C., Oliveira, L. C., Dugger, B. N., Cheung, S. C., & Chuah, C. N. (2021). Joint Semi-supervised and Active Learning for Segmentation of Gigapixel Pathology Images with Cost-Effective Labeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 591-600).

[14] Lai, Z., Wang, C., Hu, Z., Dugger, B. N., Cheung, S. C., & Chuah, C. N. (2021, November). A semi-supervised learning for segmentation of gigapixel histopathology images from brain tissues. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1920-1923). IEEE.

[15] Polson, J., Zhang, H., Nael, K., Salamon, N., Yoo, B., Kim, N., ... & Arnold, C. W. (2021, November). A Semi-Supervised Learning Framework to Leverage Proxy Information for Stroke MRI Analysis. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2258-2261). IEEE.

[16] Goschenhofer, J., Hvingelby, R., Rügamer, D., Thomas, J., Wagner, M., & Bischl, B. (2021, December). Deep semi-supervised learning for time series classification. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 422-428). IEEE.

[17] Jawed, S., Grabocka, J., & Schmidt-Thieme, L. (2020, May). Self-supervised learning for semi-supervised time series classification. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 499-511). Springer, Cham.

[18] Cheng, J. Y., Goh, H., Dogrusoz, K., Tuzel, O., & Azemi, E. (2020). Subject-aware contrastive learning for biosignals. arXiv preprint arXiv:2007.04871.

[19] Kaya, M., Binli, M. K., Ozbay, E., Yanar, H., & Mishchenko, Y. (2018). A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Scientific data*, **5**(1), 1-16.

[20] Roots, K., Muhammad, Y., & Muhammad, N. (2020). Fusion convolutional neural network for cross-subject EEG motor imagery classification. *Computers*, **9**(3), 72.

[21] Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, **20**(3), 45-50.

[22] Xu, X., & Liu, H. (2020). ECG heartbeat classification using convolutional neural networks. *IEEE Access*, 8, 8614-8619.

[23] Zhao, Z., Zhou, L., Duan, Y., Wang, L., Qi, L., & Shi, Y. (2022). DC-SSL: Addressing Mismatched Class Distribution in Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9757-9765).

A Appendix

Tables with detailed performance metrics for the experiments performed. The information of F1-scores complement the accuracy previously presented.

Table 1: PPG SL vs SSL

Framework	SL				Best SSL			
Label %	0.5%	1%	10%	100%	0.5%	1%	10%	100%
Accuracy	71.4% $\pm$ 11.0	69.86% $\pm$ 2.6	78.43% $\pm$ 1.4	80.6% $\pm$ 1.4	79.2% $\pm$ 3.36	74.3% $\pm$ 4.2	79.76% $\pm$ 2.3	79.92% $\pm$ 1.2
Normal F1	0.7 $\pm$ 0.06	0.4 $\pm$ 0.14	0.65 $\pm$ 0.13	0.73 $\pm$ 0.07	0.66 $\pm$ 0.18	0.54 $\pm$ 0.19	0.68 $\pm$ 0.12	0.74 $\pm$ 0.08
Artifact F1	0.69 $\pm$ 0.19	0.8 $\pm$ 0.04	0.84 $\pm$ 0.01	0.84 $\pm$ 0.004	0.84 $\pm$ 0.02	0.81 $\pm$ 0.03	0.84 $\pm$ 0.01	0.83 $\pm$ 0.01

Results for cross-validation evaluation of in-house PPG artifact detection dataset. Best SSL refers to FlexMatch.

Table 2: EEG SL vs SSL

Framework	SL				Best SSL			
Label %	0.5%	10%	50%	100%	0.5%	10%	50%	100%
Accuracy	59.1% $\pm$ 2.56	67.57% $\pm$ 0.2	68.67% $\pm$ 0.2	69.9% $\pm$ 1.2	61.5% $\pm$ 1.22	68.7% $\pm$ 0.85	72.2% $\pm$ 0.74	N/A
Hand F1	0.64 $\pm$ 0.02	0.67 $\pm$ 0.01	0.68 $\pm$ 0.01	0.69 $\pm$ 0.02	0.64 $\pm$ 0.01	0.69 $\pm$ 0.02	0.72 $\pm$ 0.01	N/A
Foot F1	0.52 $\pm$ 0.08	0.68 $\pm$ 0.01	0.68 $\pm$ 0.01	0.7 $\pm$ 0.01	0.58 $\pm$ 0.03	0.69 $\pm$ 0.02	0.72 $\pm$ 0.01	N/A

Results for cross-validation evaluation of EEG Motor Imagery dataset [21]. Best SSL refers to UDA.

Table 3: ECG SL vs SSL

Framework	SL				Best SSL			
Label %	3.5%	15%	50%	100%	3.5%	15%	50%	100%
Accuracy	86.1% $\pm$ 5.32	82.92% $\pm$ 9.7	85.3% $\pm$ 4.36	84.8% $\pm$ 6.54	86.3% $\pm$ 13.1	85.49% $\pm$ 5.7	82.16% $\pm$ 2.7	N/A
N F1	0.93 $\pm$ 0.04	0.9 $\pm$ 0.06	0.92 $\pm$ 2.94	0.92 $\pm$ 0.04	0.93 $\pm$ 0.03	0.92 $\pm$ 0.04	0.9 $\pm$ 0.01	N/A
VEB F1	0.45 $\pm$ 0.39	0.56 $\pm$ 0.15	0.59 $\pm$ 0.14	0.64 $\pm$ 0.17	0.43 $\pm$ 0.37	0.67 $\pm$ 0.12	0.52 $\pm$ 0.07	N/A
SVEB F1	0.03 $\pm$ 0.02	0.03 $\pm$ 0.05	0.08 $\pm$ 0.06	0.08 $\pm$ 0.04	0.02 $\pm$ 0.02	0.03 $\pm$ 0.04	0.08 $\pm$ 0.08	N/A

Results for cross-validation evaluation of ECG MIT-BIH dataset [20]. Best SSL refers to FlexMatch.